Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-020-00550-7 OPEN

Angry by design: toxic communication and technical architectures

Luke Munn[1]✉

Hate speech and toxic communication online is on the rise. Responses to this issue tend to offer technical (automated) or non-technical (human content moderation) solutions, or see hate speech as a natural product of hateful people. In contrast, this article begins by recognizing platforms as designed environments that support particular practices while discouraging others. In what ways might these design architectures be contributing to polarizing, impulsive, or antagonistic behaviors? Two platforms are examined: Facebook and YouTube. Based on engagement, Facebook's Feed drives views but also privileges incendiary content, setting up a stimulus–response loop that promotes outrage expression. YouTube's recommendation system is a key interface for content consumption, yet this same design has been criticized for leading users towards more extreme content. Across both platforms, design is central and influential, proving to be a productive lens for understanding toxic communication.

[1] Digital Cultures Institute, Auckland, Aotearoa, New Zealand. ✉email: luke.munn@gmail.com

Introduction

Hate speech online is on the rise (Oboler, 2016; Perrigo, 2019; Pachego and Melhuish, 2020)[1]. The response to this rise has broadly taken two approaches to harm reduction on platforms. The first approach is technical, attempting to develop software models to detect and remove problematic content. Indeed over the last few years in particular, significant attention has been directed at abusive speech online, with huge amounts of work poured into constructing and improving automated systems (Pavlopoulos et al., 2017; Fortuna and Nunes, 2018). Articles in computer science and software engineering in particular often claim to have studied the failings of previous techniques and discovered a new method that finally solves the issue (Delort et al., 2011; Mulla and Palave, 2016; Tulkens et al., 2016). And yet the inventiveness of users and the ambiguity of language mean that toxic communication remains complex and difficult to address. Technical understanding of this content will inevitably be limited, explains researcher Robyn Caplan (quoted in Vincent, 2019), because automated systems are being asked to understand human culture—racial histories, gender relations, power dynamics and so on—"a phenomenon too fluid and subtle to be described in simple, machine-readable rules".

The second approach is non-technical, stressing that hate speech online is a problem that only humans can address. This framing, not incorrectly, points out that automated interventions will always be inherently limited, unable to account for the nuances of particular contexts and the complexities of language. The response is to dramatically expand content moderation teams. In May 2018, for example, Facebook announced that it would be hiring 10,000 new workers into it's trust and safety team (Freeman, 2018). However, the toll for those carrying out this kind of work, where hate speech, graphic images, and racist epithets must be carefully reviewed, is incredibly high, leading to depression and other mental health issues. In being forced to parse this material, workers "do not escape unscathed" (Madrigal, 2017). As well as the hazards of the content itself, employees are often under intense pressure to meet performance targets, an anxiety that only adds to the inherent psychological toll (Newton, 2019).

In addition to these two approaches, there also seems to be a popular assumption, evidenced in online comments and in more mainstream literature, that hate speech is the natural product of hateful people. One user stated that the toxic comments she encountered online were simply produced by rude and frustrated people, perhaps with a difficult background or early life, who have not been taught general manners. Another blog post blames toxic communication on an inherently toxic individual, someone with a predilection for hating or bullying, racism or sexism (Jennings-Edquist, 2014). In this understanding, hate speech results from people translating their fundamental nastiness in the offline world into the online environment.

In contrast to the approaches and assumptions discussed above, this study adopts a design-centric approach. It seeks to understand how hate might be facilitated in particular ways by hate-inducing architectures. Just as the design of urban space influences the practices within it (Jacobs, 1992; Birenboim, 2018), the design of platforms, apps and technical environments shapes our behavior in digital space. This design is not a neutral environment that simply appears, but is instead planned, prototyped, and developed with particular intentions in mind. Indeed, a platform can be conceived as a set of "core design problems" (Tura et al., 2018, Table 1).

This method thus examines a platform's interfaces, architectures, and functionality, focusing on the types of communicative practices and social interactions they afford (Bucher and Helmond, 2017). As Gillespie (2017, n.p.) argues, these structures:

> are designed to invite and shape participation toward particular ends. This includes what kind of participation they invite and encourage; what gets displayed first or most prominently; how the platforms design navigation from content to user to exchange… and how they organize information through algorithmic sorting, privileging some content over others in opaque ways. And it includes what is not permitted, and how and why they police objectionable content and behavior.

A platform's design is the result of certain decisions, and these decisions have influence. Acknowledging this influence allows us to draw "connections between the design (technical, economic, and political) of platforms and the contours of the public discourse they host" (Gillespie, 2015, p. 2). How might the design of technical environments be promoting toxic communication?

This project examined two notable platforms: Facebook and YouTube. Both platforms have millions or even billions of monthly active users. Both platforms have a global reach, with access available in hundreds of countries worldwide. And both have been linked to hate speech, online harassment, and more overt acts of physical violence in the "real world". Both platforms are thus highly influential, shaping the beliefs and ideologies of individuals, their media production and consumption, and their relations to others on an everyday basis.

Following the method sketched above, this analysis meant identifying key elements of the platform's design—the news feed or a recommendation engine, for instance. The analysis then honed in on these architectures and affordances, asking how this design operates, what is its logic, and what type of speech and behavior does it encourage. While using these platforms provided insight, these questions frequently also meant drawing on secondary literature from designers, platform users, and software engineers. This core design analysis was supplemented by two unstructured interviews. The first was with a young social media user. The second was with a former online community manager, whose previous role ranged from guiding forum discussions to offering user assistance and moderating content. Both of these inputs are drawn on at several points to offer a "vernacular" perspective on design (McVeigh-Schultz and Baym, 2015)—foregrounding how it is perceived and dealt with on a practical everyday level.

While this method is novel in some ways, the attention to the design of platforms and their potential to shape behavior is not unprecedented. Over the last few years, we have witnessed a confessional moment from the designers of platforms.[2] Designers have admitted that their systems are addictive and exploit negative "triggers" (Lewis, 2017). They have explained that Facebook's design privileges base impulses rather than considered reflection (Bosker, 2016). Others have spoken about their tools "ripping apart the social fabric of how society works" (Vincent, 2017). And these confessions have been echoed with criticism and studies from others. Social media enables negative messages to be distributed farther and faster (Vosoughi et al., 2018) and its affordances enable anger to spread contagiously (Fan et al., 2016). The "incentive structures and social cues of algorithm-driven social media sites" amplify the anger of users over time until they "arrive at hate speech" (Fisher and Taub, 2018). In warning others of these negative social effects, designers have described themselves as canaries in the coal mine (Mac, 2019).

Indeed, we have already begun witnessing the fallout of platform-amplified hate. Shootings in El Paso, Pittsburgh, and Christchurch have been linked to users on Gab and 8chan (Mezzofiore and O'Sullivan, 2019; Silverstein, 2018). Ethnic violence against Rohingya has been connected to material circulating on Facebook (Stevenson, 2018). And anti-Muslim Tweets have been correlated with anti-Muslim hate crime (Williams et al., 2020). These overt acts of hate in the "real world" materialize this issue and highlight its significant stakes. Toxic communication is not just a nuisance or a nasty byproduct of online environments, but has more fundamental implications for human rights. "Online hate is no less harmful because it is online", stressed a recent U.N. report (Kaye, 2019): "To the contrary, online hate, with the speed and reach of its dissemination, can incite grave offline harm and nearly always aims to silence others". Hate forms a broad spectrum with extremist ideologies at one end. Online environments allow users to migrate smoothly along this spectrum, forming a kind of pipeline for radicalization (O'Callaghan et al., 2015; Munn, 2019). In this respect, the hate-based violence of the last few years is not random or anomalous, but a logical result of individuals who have spent years inhabiting hate-filled spaces where racist, sexist, and anti-Semitic views were normalized.

Very recently, then, a new wave of designers and technologists have begun thinking about how to redesign platforms to foster calmer behavior and more civil discourse. How might design create ethical platforms that enhance users wellbeing (Han, 2019)? Could technology be designed in a more humane way (Harris, 2019)? And what would be the core principles and processes of such designs (Yablonski, 2019)? Identifying a set of hate-promoting architectures would allow designers and developers to construct future platforms that mitigate communication used to harass or harm, and instead construct more inclusive and affirmative environments.

This article picks up on this nascent work, tracing the relationship between technical architectures and toxic communication. It examines two highly influential global platforms, Facebook and YouTube, unpacking the design of several key features, identifying how they are problematic, and suggesting some possible alternatives.

Platform analysis: Facebook

Facebook is the giant of social media. With 2.41 billion active users worldwide (Noyes, 2019), it is the largest platform, and arguably one of the most significant. On average, users spend 58 min every day on the platform (Molla and Wagner, 2018). While some signs indicate that the platform is plateauing in terms of use, these statistics remain compelling and mean that it cannot be overlooked. From the perspective of this project, Facebook is a technically mediated environment where vast numbers of people spend significant amounts of time. Yet if the platform is influential, it is also increasingly recognized as detrimental. "As Facebook grew, so did the hate speech, bullying and other toxic content on the platform", one investigation found (Frenkel et al., 2018), "when researchers and activists in Myanmar, India, Germany and elsewhere warned that Facebook had become an instrument of government propaganda and ethnic cleansing, the company largely ignored them". What kinds of experiences are all of these users having, and how does the design of this environment contribute to this? Rather than calm and civil, this analysis will show how the platform's affordances can induce experiences that are stressful and impulsive, establishing some of the key conditions necessary for angry communication.

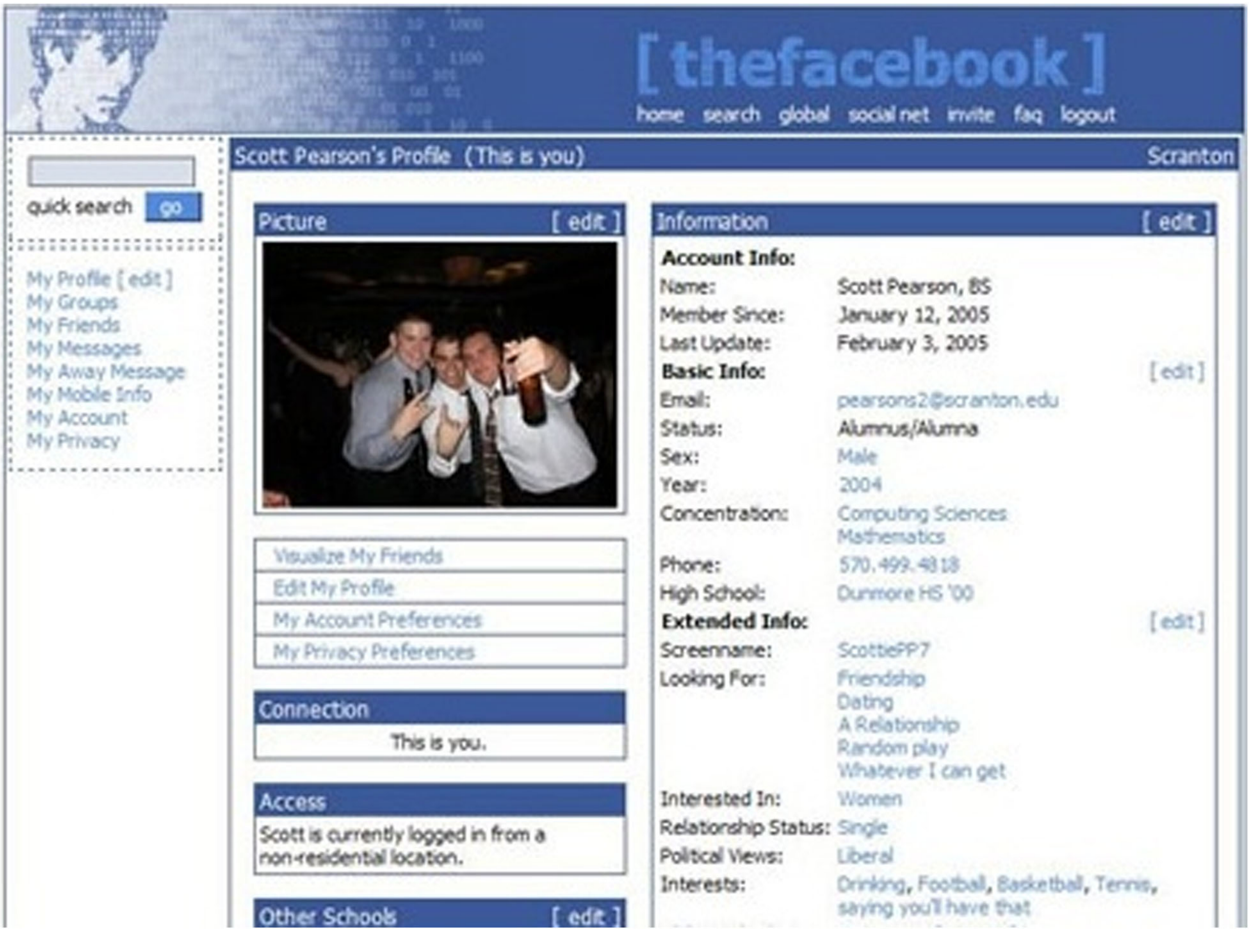
A design approach to Facebook stresses that it was designed—a result of particular decisions made over time. For users, Facebook appears as a highly mature and highly refined environment. Every area has undergone meticulous scrutiny and crafting by teams of developers and designers. This provides the environment with a degree of stability and authority, even inevitability. In this sense, giants like Facebook claim a kind of de-facto standard: this is the way our communication media operates. Yet Facebook has evolved significantly since its inception. Launching in 2004, the site was billed as an "online directory"; in these early days, the site emulated the approach of MySpace, where each user had a profile, populated with fields for status, education, hobbies, relationships, and so on; in 2007, Facebook added a Mini-Feed feature that listed recent changes to friends profiles, and in 2011 Facebook released the Timeline that "told the story of your life" as a move away from the directory or database structures of the past (Albenesius, 2014). Rather than inevitable, then, the design evolution of Facebook reminds us that it has evolved through conscious decisions in response to a particular set of priorities (Fig. 1).

Design wise, the Feed remains one of the key pieces of functionality within Facebook. The Feed, or the News Feed as it is officially known, is described by the company as a "personalized, ever-changing collection of photos, videos, links, and updates from the friends, family, businesses, and news sources you've connected to on Facebook" (Facebook, 2019). It is the first thing that users see when bringing up the app or entering the site. It is the center of the Facebook experience, the core space where content is presented to users. What's more, because user actions are primed by this content and linked to it—whether commenting on a post, sharing an event, or liking a status update—the Feed acts as the gateway for most user activity, structuring the actions they will perform during that particular session. Indeed, for many users, Facebook is the Feed and the Feed is Facebook (Manjoo, 2017).

Key to the Feed is the idea of automatic curation. Before the Feed, users would have to manually visit each one of their friend's profile pages in order to discover what had changed in his or her life. Once introduced, the Feed now carries out this onerous task for each user. "It hunts through the network, collecting every post from every connection—information that, for most Facebook users, would be too overwhelming to process themselves" (Manjoo, 2017). In this sense, the Feed provides both personalization and convenience, assembling a list of updates and bringing them together into a single location. Yet from a critical design perspective (Dunne and Raby, 2001; Dunne, 2006; Bardzell and Bardzell, 2013), this begs some fundamental questions about values, ideologies, and norms. What is prioritized in this Feed, bubbling to the top of view and clamoring for a user's attention? What is deemphasized, only appearing after a long scroll to the bottom? And what are the factors that influence this invisible curation work? In short: what is shown, what is hidden, and how is this decided (see Fig. 2)?

The Feed is designed according to a particular logic. Since 2009, stories are not sorted chronologically, where updates from friends would simply be listed in reverse order, with the most recent appearing first (Wallaroo Media, 2019). While this change induced a degree of backlash from users, the chronology itself proved to be overwhelming, especially with the hundreds of friends that each user has. "If you have 1500 or 3000 items a day, then the chronological feed is actually just the items you can be bothered to scroll through before giving up", explains analyst Benedict Evans (2018), "which can only be 10% or 20% of what's actually there". Instead, the Feed is driven by Engagement. In this design, Facebook weighs dozens of factors, from who posted the content to their frequency of posts and the average time spent on this piece of content. Posts with higher engagement scores are included and prioritized; posts with lower scores are buried or excluded altogether (see Fig. 3).

The problem with such sorting, of course, is that incendiary, polarizing posts consistently achieve high engagement (Levy, 2020, p. 627). This content is meant to draw engagement, to provoke a reaction. Indeed, in 2018 an internal research team at



[thefacebook]

home search global social net invite faq logout

quick search go

My Profile [edit]
My Groups
My Friends
My Messages
My Away Message
My Mobile Info
My Account
My Privacy

Scott Pearson's Profile (This is you) Scranton

Picture [edit]

Visualize My Friends
Edit My Profile
My Account Preferences
My Privacy Preferences

Connection

This is you.

Access

Scott is currently logged in from a non-residential location.

Other Schools [edit]

Information [edit]

Account Info:

Name:	Scott Pearson, BS
Member Since:	January 12, 2005
Last Update:	February 3, 2005

Basic Info: [edit]

Email:	pearsons2@scranton.edu
Status:	Alumnus/Alumna
Sex:	Male
Year:	2004
Concentration:	Computing Sciences Mathematics
Phone:	570.499.4818
High School:	Dunmore HS '00

Extended Info: [edit]

Screenname:	ScottieP7
Looking For:	Friendship Dating A Relationship Random play Whatever I can get
Interested In:	Women
Relationship Status:	Single
Political Views:	Liberal
Interests:	Drinking, Football, Basketball, Tennis, saying you'll have that

Fig. 1 Early Facebook Screenshot. Early screenshot from "The Facebook" indicating its significant design progression over time.

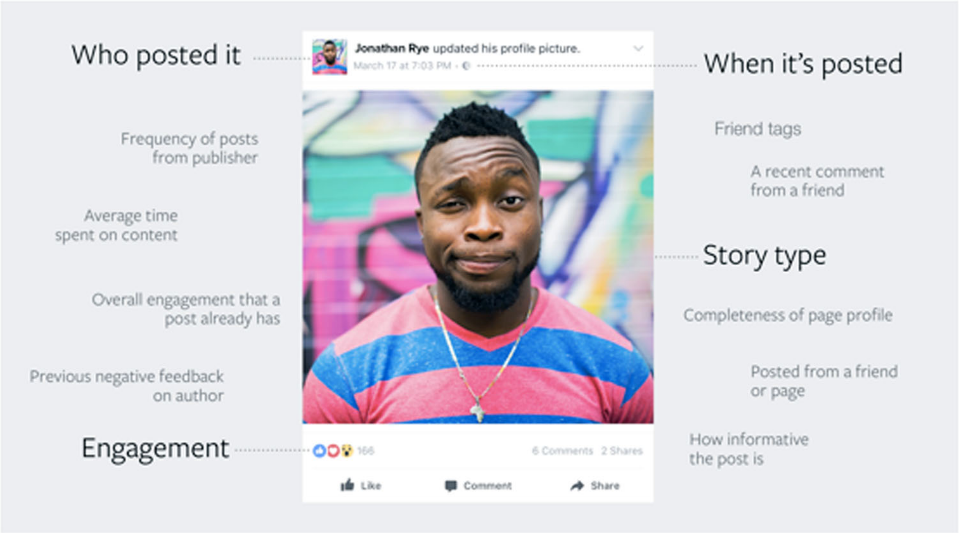

Who posted it
Jonathan Rye updated his profile picture.
March 17 at 7:03 PM
When it's posted
Frequency of posts from publisher
Friend tags
A recent comment from a friend
Average time spent on content
Story type
Overall engagement that a post already has
Completeness of page profile
Previous negative feedback on author
Posted from a friend or page
Engagement
166
6 Comments 2 Shares
How informative the post is
Like
Comment
Share

Fig. 2 News Criteria. Screenshot of Facebook page listing some of the criteria used by its News Feed.

Facebook reported precisely this finding: by design it was feeding people "more and more divisive content in an effort to gain user attention and increase time on the platform" (Horwitz and Seetharaman, 2020). However, Facebook management ignored these findings and shelved the research.

This divisive material often has a strong moral charge. It takes a controversial topic and establishes two sharply opposed camps, championing one group while condemning the other. These are the headlines and imagery that leap out at a user as they scroll past, forcing them to come to a halt. This offensive material hits a nerve, inducing a feeling of disgust or outrage. "Emotional reactions like outrage are strong indicators of engagement", observes designer and technologist Tobias Rose-Stockwell (2018), "this kind of divisive content will be shown first, because it

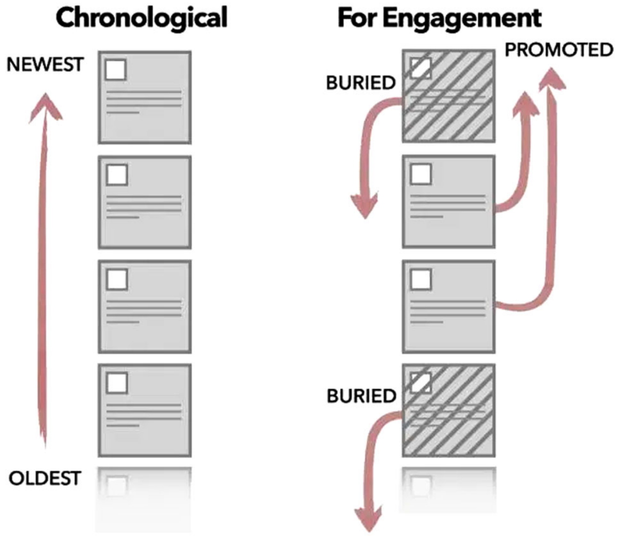


Fig. 3 Content Prioritization. Diagram from Rose-Stockwell showing the change in content prioritization (reproduced with permission).

captures more attention than other types of content". While speculative, perhaps sharing this content is a way to offload these feelings, to remove their burden on us individually by spreading them across our social network and gaining some sympathy or solidarity.

The design of Facebook means that this forwarding and redistribution is only a few clicks away. As the user I interviewed stated: "it is so easy to share stuff". Moreover, the networked nature of social media amplifies this single response, distributing it to hundreds of friends and acquaintances. They too receive this incendiary content and they too share, inducing what Rose-Stockwell (2018) calls "outrage cascades — viral explosions of moral judgment and disgust". Outrage does not just remain constrained to a single user, but proliferates, spilling out to provoke other users and appear in other online environments.

At its worst, then, Facebook's Feed stimulates the user with outrage-inducing content while also enabling its seamless sharing, allowing such content to rapidly proliferate across the network. In increasing the prevalence of such content and making it easier to share, it becomes normalized. Outrage retains its ability to provoke engagement, but in many ways becomes an established aspect of the environment. For neuroscientist Molly Crockett, this is one of the keys to understanding the rise of hate speech online. Crockett (2017, p. 770) stresses that "when outrage expression moves online it becomes more readily available, requires less effort, and is reinforced on a schedule that maximizes the likelihood of future outrage expression in ways that might divorce the feeling of outrage from its behavioral expression". Design, in this sense, works to reduce the barrier to outrage expression. Sharing a divisive post to an audience of hundreds or thousands is just a click away.

How might the Feed be redesigned? Essentially there are two separate design problems here. Firstly, there is the stimulus aspect —the content included in the Feed. While the Feed's filtering operations undoubtedly remain highly technical, its logics can be understood through a design decision to elevate and amplify "engaging" content. Facebook has admitted that hate speech is a problem and has redesigned the Feed dozens of times since its debut in an effort to curtail this problem and the broader kind of misinformation that often stirs it up (Wallaroo Media, 2019). But the core logic of engagement remains baked into the design of the Feed at a deep level. Design, then, might start by experimenting quite concretely with different kinds of values. If the hyperlocal was privileged, for example, then posts from friends or community members in a 5 km radius might only be shown. This would be more mundane in many ways—everyday updates from those in our immediate vicinity rather than vicious attacks from anyone in a friend network. Or following the success of more targeted messaging apps like Messenger and WhatsApp, the Feed might emphasize close familial or friend connections above all. This pivot to a more intimate relational sphere would certainly be quieter and less "engaging" but ultimately more meaningful and civil.

Secondly, there is the response aspect—the platform affordances that make outrage expression online more effortless. Such expression is often impulsive, done in the moment, and so one possible design focus would be time itself. Temporality is a key part of community, stated the community manager I interviewed. "Legacy environments" such as traditional forums simply moved slower, she recalled, and in general there was "just more oxygen between things happening". This time gap between reading and posting provided both a kind of deceleration and de-escalation, a chance to pause and reconsider. Rather than an instant reaction, would a built-in delay add a kind of emotional weight to such an action? An interval of a few seconds, even if nominal, might introduce a micro-reflection and suggest an alternative response. As a means of combating the effortless and abstract nature of outrage expression, Rose-Stockwell (2018) suggests a number of humanizing prompts that might be designed into platforms: an "empathetic prompt" that asks whether a user really wants to post hurtful content; an "ideological prompt" that stresses how this post will never be seen by those with opposing viewpoints; and a "public/private prompt" that would allow disagreements to take place between individuals rather than in the pressurized public arena. Such design interventions, while clearly not silver bullet solutions, might contribute in their own small way towards a more civil and less reactive online environment.

Platform analysis: YouTube

YouTube remains a juggernaut of online spaces. Recently, it crossed the threshold of 2 billion logged-in users per month (Saima, 2019). Perhaps even more important for this research project is the time spent by users within this environment. Users spend around 250 million hours on the video sharing platform every day (Saima, 2019). The time "inhabiting" YouTube marks it out as distinct from Facebook, and suggests a different kind of influence over time, something slower and more subtle. Indeed, as will be discussed, radicalized individuals have noted how influential YouTube was in shifting their worldview over longer periods of time, a medial pathway that nudged them towards an angrier and more extremist stance (Roose, 2019). While this is just one highly politicized facet of YouTube, it signals the stakes involved here—not only the anger available to be tapped into, but the influence such an environment might have in shaping the ideologies of its vast population.

One key focus of recent critiques of YouTube has been its recommendation engine (Regner, 2014; Schmitt et al., 2018; Ribeiro et al., 2020). The design of the recommendation system is central to YouTube's user experience for two reasons. Firstly, it determines the content of each user's homepage. Upon arriving on the site, each user is presented with rows of recommended videos, with each row representing an interest (e.g. gaming), channel (e.g. the Joe Rogan Experience), or an affiliation ("users who watched X enjoyed Y"). As with similar designs such as Netflix, the YouTube homepage is the first thing that users interact with, and the primary "jumping off" point for determining what to watch.

Secondly, the YouTube recommendation system is crucial because it also determines the related videos appearing in the sidebar next to the currently playing video. By default, the Autoplay feature is on, meaning that these sidebar videos are queued to play automatically after the current video. This design feature means that, even if the user does nothing further, the next video in this queue will play. Even if the Autoplay feature has been manually turned off, this sidebar, with its dozens of large thumbnails, presents the most obvious gateway to further content. With a single click, a user can move onto a video which is related to the one they are currently viewing.

From a design perspective, the homepage and the sidebar form the crucial interfaces into content consumption. Search, while possible, is a manual process that requires more effort and has been deemphasized. Browsing recommended results, with its scrolling and tapping, provides a more frictionless user experience. It is unsurprising then, that "we're now seeing more browsing than searching behavior", stated one YouTube designer (Lewandowski, 2018), "people are choosing to do less work and let us serve them". This shift has meant an even greater role for the recommendation engine. In theory, users can watch any video on the vast platform; in practice, they are encouraged towards a very specific subset of content. Indeed, YouTube's Chief Product Officer revealed that recommended videos account for over 70% of watching time on the platform (Solsman, 2018). This is a single algorithmic system that exerts enormous force in determining what kinds of content users are exposed to and what paths they are steered down.

How is this recommendation system designed? In a paper on its high-level workings, YouTube engineers explain that it comprises two stages. In the first stage, "the enormous YouTube corpus is winnowed down to hundreds of videos" that are termed candidates (Covington et al., 2016, p. 192). These candidates are then ranked by a second neural network, and the highest ranked videos presented to the user. In this way, the engineers can be "certain that the small number of videos appearing on the device are personalized and engaging for the user" (Covington et al., 2016, p. 192). Based on hundreds of signals, users are presented with content that is attractive by design: hooking into their interests, goals, and beliefs. This recommendation engine is not static, but rather highly dynamic and updated in real-time. Your profile incorporates your history, but also whatever you just watched. As YouTube's engineers (Covington et al., 2016, p. 191) explain, it must be "responsive enough to model newly uploaded content as well as the latest actions taken by the user". As content is consumed, an individual's interests and ideologies in turn are shaped (Fig. 4).

Of course, these technical explanations remain at a high-level. The recommendation system, as a proprietary technology owned and operated by YouTube, will always remain to some extent a black box. Yet even these general principles provide insight into the system's design. First, the system is designed to promote "engaging" videos. Which videos are most engaging? As one former developer (Chaslot, 2019) explains:

> We know that misinformation, rumors, and salacious or divisive content drives significant engagement. Even if a user notices the deceptive nature of the content and flags it, that often happens only after they've engaged with it. By then, it's too late; they have given a positive signal to the algorithm. Now that this content has been favored in some way, it gets boosted, which causes creators to upload more of it. Driven by AI algorithms incentivized to reinforce traits that are positive for engagement, more of that content filters into the recommendation systems. Moreover, as soon as the AI learns how it engaged one person, it can reproduce the same mechanism on thousands of users.

Recommending content based on engagement, then, often means promoting incendiary, controversial, or polarizing content. The closer a video gets to the edge of what's allowed under YouTube's policy, the more engagement it gets (Maack, 2019). In other words, as even Zuckerberg (2018) has admitted, borderline content is more engaging. Because of this dynamic, designing for engagement goes beyond mere customer satisfaction to deeply influence the kind of content that promoted. As the developer quote above suggests, the system's design establishes a series of powerful feedback loops. Creators create more of this toxic yet high-performing content and the system recommends it more often to users, not only individually, but at scale.

Secondly, the system is designed to be responsive, to be dynamic enough to generate new recommendations based on what was last viewed. The design challenge, as the engineers explain (Covington et al., 2016, p. 194), is to predict "the next watched video". While again high-level, this creates a design with a degree of self-similarity, promoting more of the same kind of content. And yet if this content stays within the same topic, it is typically more intense, more extreme. "However extreme your views, you're never hardcore enough for YouTube" attests one article (Naughton, 2018). Based on the strong performance of borderline content discussed earlier, YouTube's recommendations often move from mainstream content to more incendiary media, or politically from more centrist views to right and even far-right ideologies.

The dynamism designed into the recommendation system establishes a vector, a gradual movement as each video is completed. Based on the current values designed into the system, users can be suggested material that progressively becomes more controversial, more political, more outrage-inducing, and in some cases, more explicitly racist, sexist, or xenophobic (O'Callaghan et al., 2015). Indeed, one analysis (Munn, 2019) suggests that YouTube can form a key part of an "alt-right pipeline": users are incrementally nudged down a medial pathway towards more far-right content, from anti-SJW videos which demean so-called "social justice warriors" to gaming related misogyny, conspiracy theories, the white supremacism of "racial realism" and thinly veiled anti-Semitism. In a recent paper analyzing approximately 330,925 videos across 349 channels, a study found that "users consistently migrate from milder to more extreme content", shifting from viewing so-called Alt-Lite material to more strident Alt-Right channels (Ribeiro et al., 2020, p. 131).

What is particularly powerful about this design is its automatic and step-wise quality. Users do not consciously have to select the next video, nor jump suddenly into extreme material. Instead, there is a slow progression, allowing users to acclimate to these views before smoothly progressing onto the next step into their

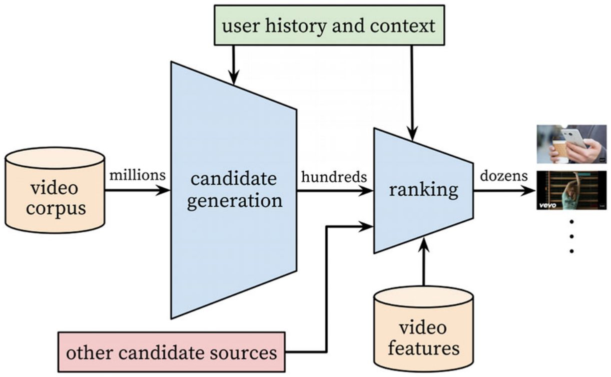
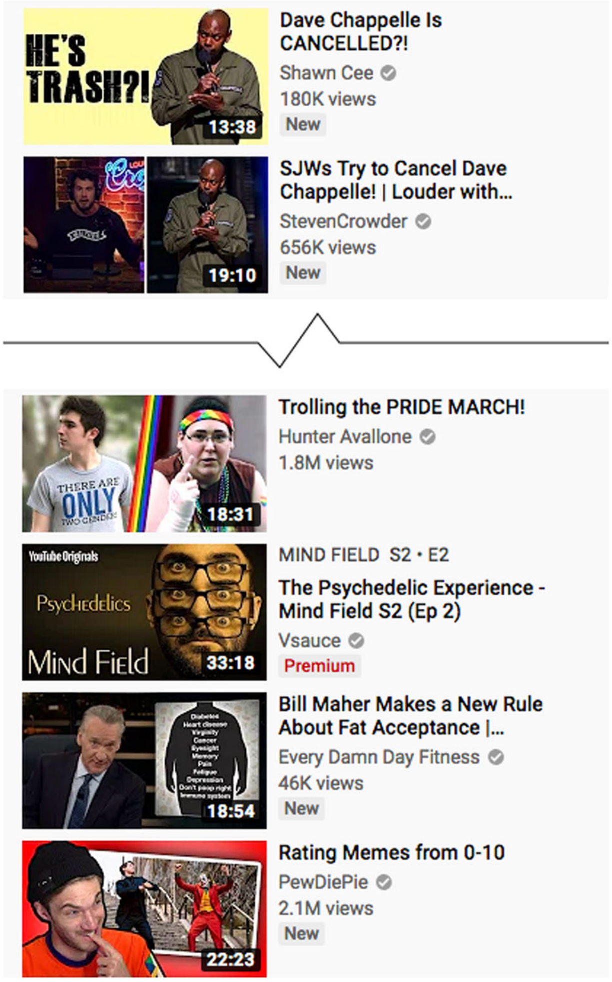


Fig. 4 Recommendation System. Diagram from YouTube engineers indicating how the recommendation system works (reproduced with permission).

journey. Recommendations are “the computational exploitation of a natural human desire: to look ‘behind the curtain’, to dig deeper into something that engages us”, observes sociologist Zeynep Tufekci (2018): “As we click and click, we are carried along by the exciting sensation of uncovering more secrets and deeper truths. YouTube leads viewers down a rabbit hole of extremism, while Google racks up the ad sales”. At the far end of this journey is an angry and radicalized individual, a figure that has increasingly emerged over the last few years, from Christchurch in New Zealand to El Paso, Texas and Poway, California in the United States. Yet along with these extreme examples, equally troubling is the thought of a broader, more unseen population of users who are gradually being exposed to more hateful material.

The result of these design choices is that the recommendation system emerges as a hate-inducing architecture. From a metrics point of view, the system is successful, delivering “engaging” content while ramping up view counts and watch time on the platform. And yet to do so, the system appears to consistently suggest divisive, untrue, or generally incendiary content. “YouTube drives people to the Internet’s darkest corners” warns one article (Nicas, 2018). In this sense, the design of the current recommendation serves the company well, but not necessarily individual users or online communities, particularly those that are already marginalized (Fig. 5).



Dave Chappelle Is CANCELLED?!
Shawn Cee
180K views
New

SJWs Try to Cancel Dave Chappelle! | Louder with...
StevenCrowder
656K views
New

Trolling the PRIDE MARCH!
Hunter Avallone
1.8M views

MIND FIELD S2 • E2
The Psychedelic Experience - Mind Field S2 (Ep 2)
Vsauce
Premium

Bill Maher Makes a New Rule About Fat Acceptance |...
Every Damn Day Fitness
46K views
New

Rating Memes from 0-10
PewDiePie
2.1M views
New

Fig. 5 YouTube Recommendations. Screenshots showing anti-SJW (social justice warrior) and anti-LGBTQ+ recommendations in response to viewing a centrist-right video by popular talk show host Joe Rogan.

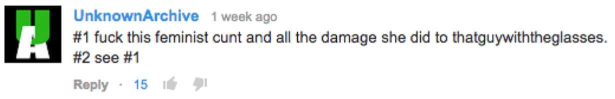
It should be noted that one recent working paper has questioned the role of the recommendation system in hate speech and far-right indoctrination. Munger and Phillips (2019) argue that the central role given to the recommendation engine is overplayed, and suggests instead a supply and demand explanation. For the duo, YouTube lowers the barriers of media production to almost zero, it offers easy distribution online through hosting and sharing, and it incentivizes content creation via monetization. These conditions have led to a diversification of channels that politically stretch beyond the mainstream center-left/center-right poles. As Munger and Phillips argue (2019, p. 6): “these aspects of YouTube allow new communities that cater increasingly well to audiences’ ideas to form”. The YouTube platform allows for the proliferation of niche media and a greater variety of alt-right and far-right material. The duo essentially argue that a radicalized audience already existed, it was simply constrained by too little supply of radical material.

On the one hand, the report is a productive reminder that social media is a *socio*technical system. Technologies are never purely determinist and any analysis should strive to account for the political and cultural background of users, their relations to others in the world, and the racial and gendered worldviews that “link” content together, even without an engine or automated system. As Rebecca Lewis (2018) has shown, the network of alt-right influencers on YouTube is a social network in the conventional sense—a web of individuals who share particular ideologies, use common phrases, and even recommend each other’s channels organically through formats like the talk show.

On the other hand, however, Munger and Phillips are using a rather conventional economic model to understand online environments. Their analysis presupposes an offline, radicalized audience with their minds already made up. In doing so, it fails to register the psychological and cognitive force exerted by platform environments, a force potentially magnified both by time spent consuming media and by the young age of particular users. Contrary to the duo’s straw man caricature of such influence as a “zombie bite”, this force is not an instant contagion, but something far more drawn out and subtle, a quiet influence that alters individuals as they inhabit online spaces over the months and years. As Wendy Chun (2017, p. x) observes, media exerts force over a “creepier, slower, more unnerving time”, effectively “disappearing from consciousness”. Media derives its power precisely by catering to the curiosities and desires of the user rather than overpowering them.

Along with the recommendation engine, another problematic design element identified in this analysis is YouTube’s comment system. For years, YouTube has consistently held a reputation for being an environment with some of the most toxic and vitriolic comments online (Tait, 2016). Even those used to online antagonism admitted that “you will see racist, sexist, homophobic, ignorant, and/or horrible comments on virtually every popular post” and yet the same post from 2013 naively claims that the problem will soon be solved with new technical features (Rose, 2013). Far from being solved, the years since have seen toxic communication on the platform proliferate and take on concerning new forms. While regarded as a “cesspool” for over a decade, the latest indictment has been a large number of predatory and sexual comments on the videos of minors (Alexander, 2018).

Why is YouTube so toxic, so angry? One common explanation is that YouTube is simply one of the largest platforms. For some, its extremely broad demographic explains its trend towards the lowest common denominator in terms of intelligent, relevant commentary. Yet while the platform certainly has a massive userbase, there also seem to be clear design decisions exacerbating these toxic comments. “Comments are surely affected by *who*



UnknownArchive 1 week ago

#1 fuck this feminist cunt and all the damage she did to thatguywiththeglasses.
#2 see #1

Reply · 15

Fig. 6 Toxic Comment. Screenshot showing just one example from the many toxic comments on YouTube.

writes them", admits one analysis (Polymatter, 2016), "but *how* a comment system is designed greatly affects *what* is written". For instance, YouTube comments can be upvoted or downvoted, but downvoting doesn't lower the number of upvotes. This suggests a design logic that favors any kind of engagement, whether positive or negative. The result is that provocative, controversial, or generally polarizing comments seem to appear towards the top of the page on every video (Fig. 6).

The design choices built into both YouTube's recommendation engine and its comment system might be understood as natural outcomes of an overarching set of company values. As recent articles have shown (Bergen, 2019), YouTube as purposefully ignored warnings of its toxicity for years—even from its own employees—in its pursuit of one value: engagement. Of course, this should come as no surprise for a publicly listed company driven by shareholder values and the broader dictates of capitalism. However it opens the question into what values are prioritized within online environments and how design supports them. Rather than grand vision statements or aspirational company values, what are the incentives built into platforms at the level of design: features, metrics, interfaces, and affordances?

Echoing this low-level design influence, the community manager I interviewed underlined how the typical all-consuming focus on likes and shares could be damaging. A key part of a community manager's role is to foster healthy relations between members, to encourage beneficial content, and to block, delete or demote toxic posts—in short, to facilitate "more of the good and less of the corrosive". But her fellow community managers often speak of "algorithm chasing", where they attempt to combat or counteract the features built into the systems they use. There are often "competing logics" on a platform, she explained, an opposition between the value of creating a cohesive and civil community, and the values seen as necessary for platform growth and revenue such as expanding a user-base, extending use times, and attracting advertisers. Social media and community are often an awkward fit, and "marketing efficiencies are not social efficiencies". On YouTube specifically, these designs privilege engagement above all else, resulting in a community that can be toxic and angry. Yet design might be rethought to prioritize an alternative set of values.

How might design contribute to a calmer, more considerate and more inclusive environment? One concrete intervention would be a redesigned recommendation system. Programmer and activist Francis Irving (2018) has found that the current system described earlier is both populist, prioritizing the popular, and short-term, using criteria to find videos that you'll watch the longest. What kind of design interventions would make it more conducive to user well-being? For one, the system could be intentionally broadened, breaking its hyper-focused bubble and instead providing access points into a range of communities and a diversity of political views—even those that run counter to the user. Of course, other possibilities abound. Irving (2018) suggests one playful alternative: ask whether a YouTube user is more or less happy 6 months later, and use this signal as a way to improve video recommendations. As another option, Irving (2018) speculates about removing automated recommendations altogether, and moving to a more user-centered recommendation model. Like film or music, such a model would elevate taste makers who could curate great "playlists" of content.

Secondly, the comment system might be rethought entirely. It is clear that the current upvote/downvote binary is not working, rewarding quick immediate comments that are provocative—at best flippant, at worst, hateful or degrading. It also seems apparent that the relative anonymity of commenters and lack of any concept of reputation means that there is no real disincentive for consistently generating toxic comments. As one analysis noted (Polymatter, 2016): "Each comment stands on its own, attached to nothing, bringing out the worst in every commenter". Introducing a reputation system into this environment would be one concrete design intervention. Reddit, for example, features a Karma system that rewards high quality comments while docking points for comments against community guidelines. Such a system, while naturally not perfect, significantly "thickens" the identity of a user. Each user has a history of contributions and comments that persists over time. Based on this past behavior, they have a combined score that signals whether or not the community has found their contributions helpful or beneficial. Even if this score is mainly symbolic, these reputation systems hook into offline conventions of social standing within a community, introducing a degree of accountability.

Conclusion

This article has asked how design might be contributing to polarizing, impulsive, or antagonistic behaviors. After selecting two global platforms, it approached the problem of online hate from a design perspective, identifying key affordances and structures, investigating how they function, and showing how they facilitate particular practices while discouraging others.

Based on engagement, Facebook's Feed drives clicks and views, but also privileges incendiary content, setting up a stimulus–response loop where outrage expression becomes easier and even normalized. Alternative ways of prioritizing content should be explored to decrease this kind of stimuli and in general to de-escalate the user experience, providing a slower, calmer and more civil environment. In terms of user responses to this content, design interventions might be used to question, delay, or limit the scope of hateful comments. YouTube's recommendation system is at the heart of the platform's design, exerting enormous influence on viewing and consumption. The system's design also privileges engagement, creating an environment criticized for leading users towards more extreme content. Both this recommendation system and YouTube's infamous comment system need to be thoroughly redesigned, with the section laying out several suggestions.

How feasible are such suggestions? Would these platforms realistically ever be redesigned? The prime directive of engagement, for example, is driven by monetization. It befits a corporation aiming to accelerate growth, stimulate ad revenue, and generate profits for its shareholders. After all, these platforms are a new "space of accumulation" (Fuchs, 2011), with a business model predicated on the production and extraction of data as a form of capital (Sadowski, 2019). And yet even from a purely economic perspective, engagement at any cost has been criticized. This incentive, designed deeply into the platform's interfaces and affordances, seems to encourage profiting from hate speech and other toxic communication, with both users and advertisers leaving the platform as a result (Hern, 2020). This suggests that companies like Facebook or Google—or the future platforms that will follow them—might also be searching for alternate ways of designing their products and services.

Regardless of the likelihood of a redesign in the present, one strength of a design-focused approach is that it reminds us that redesign is possible. Despite their maturity, these objects are not fixed but fluid. Each platform is the result of a set of a careful set of decisions over time. Each design element had to be conceived, prototyped, coded, tested, and launched. And what has been made can be remade. In this way, design alerts us to alternatives, to other ways of keeping us informed, structuring sociality, and valuing the people and things surrounding us. It allows us to imagine a post-Facebook/post-YouTube media environment with a different set of imperatives. Design gives us permission to speculate, to ask "what if?" (Dunne and Raby, 2014, p. 141). When our dominant technical systems seem so given, this ability to speculate about other designs becomes increasingly important.

A design approach also highlights its influence on platforms. Design privileges certain forms of content, it enables particular kinds of relations, and encourages specific forms of participation. For this reason, design proved to be a productive lens for understanding toxic communication. Of course, this study also had its limits. In particular, the degree to which design may influence individuals—and how that influence might be modulated by age, gender, class, or culture—has yet to be precisely determined. One path for research future would be to take up this challenge, producing a more quantitative analysis of design influence. Another path would be to apply this approach to other platforms: Reddit, TikTok, 4chan, and so on. Yet if this single study inevitably has constraints, it reaffirms the key role that design plays within online environments. As everyday life increasingly migrates online, platforms become crucial mediators for communication and key environments for inhabitation. These are spaces where time is spent, identities are forged, and ideologies are shaped. Understanding how these spaces might be redesigned in order to discourage hate speech and encourage civility and inclusivity remains an urgently needed task.

Received: 22 December 2019; Accepted: 16 July 2020;
Published online: 30 July 2020

Notes

[1] Because of its engagement with hate speech, toxic communication online, and far-right cultures in particular, this report features terms and language that is hateful—racist, sexist, homophobic, xenophobic, or otherwise non-inclusive. Citing such language is important to demonstrate the kinds of discourse circulating in these spaces, and indeed the degree to which in some ways they become normalized. However, it should be stressed that such language in no way reflects the views of the author or of the publisher.

[2] By "designer", I mean anyone who alters the way a platform operates, making software developers and user interface designers equally designers.

References

Albanesius C (2014) 10 years later: Facebook's design evolution. PCMag, Australia, February 4, 2014. https://au.pcmag.com/internet-2/12249/10-years-later-facebooks-design-evolution

Alexander J (2018) Can YouTube fix its comment section? Polygon. February 16, 2018. https://www.polygon.com/2018/2/16/17020326/nikolas-cruz-youtube-comment-section

Bardzell J, Bardzell S (2013) What is "critical" about critical design? In Mackay W (Ed.), Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 3297–3306). Association for Computing Machinery. https://doi.org/10.1145/2470654.2466451

Bergen M (2019) YouTube executives ignored warnings, letting toxic videos run rampant. Bloomberg.Com, April 2, 2019. https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant

Birenboim A (2018) The influence of urban environments on our subjective momentary experiences Environ Plan B 45(5):915–932. https://doi.org/10.1177/2399808317690149

Bosker B (2016) The binge breaker. Atlantic https://www.theatlantic.com/magazine/archive/2016/11/the-binge-breaker/501122/

Bucher T, Helmond A (2017) The affordances of social media platforms. In: Burgess J, Marwick A, Poell T (Eds), The Sage handbook of social media. Sage, London

Chaslot G (2019) The toxic potential of YouTube's feedback loop. Wired, July 13, 2019. https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/

Chun WHK (2017) Updating to remain the same: habitual new media. MIT Press, Cambridge

Covington P, Adams J, Sargin E (2016) Deep neural networks for YouTube recommendations. In: Proceedings of the 10th ACM conference on recommender systems—RecSys '16. ACM Press, Boston, pp. 191–198

Crockett MJ (2017) Moral outrage in the digital age. Nat Hum Behav 1 (11):769–771. https://doi.org/10.1038/s41562-017-0213-3

Crockett M (2017) How social media makes us angry all the time. Big Think. https://www.youtube.com/watch?v=fE_QoebLUFQ

Delort J-Y, Arunasalam B, Paris C (2011) Automatic moderation of online discussion sites. Int J Electron Commer 15(3):9–30. https://doi.org/10.2753/JEC1086-4415150302

Dunne A (2006) Hertzian Tales: electronic products, aesthetic experience, and critical design. MIT Press, Cambridge

Dunne A, Raby F (2001) Design Noir: the secret life of electronic objects. Springer Science Business Media

Dunne A, Raby F (2014) Speculative everything: design, fiction, and social dreaming. MIT Press, Cambridge

Evans B (2018) The death of the newsfeed. Benedict Evans. February 4, 2018. https://www.ben-evans.com/benedictevans/2018/4/2/the-death-of-the-newsfeed

Facebook (2019) News feed. News feed | Facebook media. https://www.facebook.com/facebookmedia/solutions/news-feed

Fan R, Xu KE, Zhao J (2016) Higher contagion and weaker ties mean anger spreads faster than joy in social media. Preprint at http://arxiv.org/abs/1608.03656

Fisher M, Taub A (2018) How everyday social media users become real-world extremists. N Y Times, October 10, 2018, sec. World. https://www.nytimes.com/2018/04/25/world/asia/facebook-extremism.html

Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. ACM Comput Surv 51(4):1–30

Freeman J (2018) Facebook's 10,000 new editors. Wall Str J sec. Opinion. https://www.wsj.com/articles/facebooks-10-000-new-editors-1526491169

Frenkel S, Confessore N, Kang C, Rosenberg M, Nicas J (2018) Delay, deny and deflect: how Facebook's leaders fought through crisis. N Y Times, November 14, 2018, sec. Technology. https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html

Fuchs C (2011) The Contemporary World Wide Web: Social Medium or New Space of Accumulation? In Winseck D, Jin DY (Eds), The Political Economies of Media: The Transformation of the Global Media Industries (pp. 201–220). Bloomsbury Academic

Gillespie T (2015) Platforms intervene. Soc Media+Soc 1(1):205630511558047. https://doi.org/10.1177/2056305115580479

Gillespie T (2017) Regulation of and by platforms. In: Burgess Jean, Marwick Alice, Poell Thomas (Eds), The Sage handbook of social media. Sage, London

Han L (2019) Designing for tomorrow—a discussion on ethical design. Spotify Design. January 18, 2019. https://spotify.design/articles/2019-01-18/designing-for-tomorrow-a-discussion-on-ethical-design/

Harris T (2019) Humane: a new agenda for tech. Center For Humane Technology. April 23, 2019. https://humanetech.com/newagenda/

Hern A (2020) Third of advertisers may boycott Facebook in hate speech revolt. The Guardian, June 30, 2020, sec. Technology. https://www.theguardian.com/technology/2020/jun/30/third-of-advertisers-may-boycott-facebook-in-hate-speech-revolt

Horwitz J, Seetharaman D (2020) Facebook executives shut down efforts to make the site less divisive. Wall Str J. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

Irving F (2018) Brainstorming a better YouTube recommendation algorithm. https://www.flourish.org/2018/10/brainstorming-a-better-youtube-recommendation-algorithm/

Jacobs J (1992) The death and life of great American cities, Reissue edn. Vintage, New York

Jennings-Edquist G (2014) Abusive text messages and mobile harassment are on the rise. Mamamia. November 22, 2014. https://www.mamamia.com.au/abusive-text-messages/

Kaye D (2019) Governments and internet companies fail to meet challenges of online hate—UN Expert. OHCHR. October 9, 2019. https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25174&LangID=E

Levy S (2020) Facebook: The Inside Story. Penguin UK

Lewandowski J (2018) 5 questions for YouTube's lead UX researcher interview by Amy Avery. https://www.thinkwithgoogle.com/advertising-channels/video/youtube-user-behavior-research/

Lewis P (2017) 'Our minds can be hijacked': the tech insiders who fear a smartphone dystopia. The Guardian, October 6, 2017, sec. Technology. https://www.theguardian.com/technology/2017/oct/05/smartphone-addiction-silicon-valley-dystopia

Lewis R (2018) Alternative influence: broadcasting the reactionary right on YouTube. Data & Society, New York

Maack MM (2019) 'YouTube recommendations are toxic', says Dev who worked on the algorithm. The Next Web. June 14, 2019. https://thenextweb.com/google/2019/06/14/youtube-recommendations-toxic-algorithm-google-ai/

Mac T (2019) Canary in a coal mine: how tech provides platforms for hate. A List Apart (blog). March 19, 2019. https://alistapart.com/article/canary-in-a-coal-mine-how-tech-provides-platforms-for-hate/

Madrigal AC (2017) 'The basic grossness of humans'. The Atlantic, December 15, 2017. https://www.theatlantic.com/technology/archive/2017/12/the-basic-grossness-of-humans/548330/

Manjoo F (2017) Can Facebook fix its own worst bug? N Y Times, April 25, 2017, sec. Magazine. https://www.nytimes.com/2017/04/25/magazine/can-facebook-fix-its-own-worst-bug.html

Martineau P (2019) Maybe it's not YouTube's algorithm that radicalizes people. Wired, October 23, 2019. https://www.wired.com/story/not-youtubes-algorithm-radicalizes-people/

McVeigh-Schultz J, Baym NK (2015) Thinking of you: vernacular affordance in the context of the microsocial relationship app, couple. Soc Media+Soc 1 (2):2056305115604649. https://doi.org/10.1177/2056305115604649

Mezzofiore G, O'Sullivan D (2019) El Paso shooting is at least the third atrocity linked to 8chan this year. CNN. https://www.cnn.com/2019/08/04/business/el-paso-shooting-8chan-biz/index.html

Molla R, Wagner K (2018) People spend almost as much time on instagram as they do on Facebook. Vox. June 25, 2018. https://www.vox.com/2018/6/25/17501224/instagram-facebook-snapchat-time-spent-growth-data

Morris D (2018) Facebook accused of ignoring government warnings before mob violence in Sri Lanka. Fortune, April 22, 2018. https://fortune.com/2018/04/22/facebook-ignored-sri-lanka-hate-speech/

Mulla S, Palave A (2016) Moderation technique for sexually explicit content. In: Proceedings of the 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), pp. 56–60. https://doi.org/10.1109/ICACDOT.2016.7877551

Munger K, Phillips J (2019) A supply and demand framework for YouTube politics. Penn State, University Park

Munn L (2019) Alt-right pipeline: individual journeys to extremism online. First Monday 24 (6). https://doi.org/10.5210/fm.v24i6.10108

Naughton J (2018) However extreme your views, you're never hardcore enough for YouTube. The Guardian, September 23, 2018. https://www.theguardian.com/commentisfree/2018/sep/23/how-youtube-takes-you-to-extremes-when-it-comes-to-major-news-events

Newton C (2019) Three Facebook moderators break their NDAs to expose a company in crisis. The Verge. June 19, 2019. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa

Nicas J (2018) How YouTube drives people to the internet's darkest corners. Wall Str J. February 7, 2018. https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478

Noyes D (2019) Top 20 Facebook statistics. Zephoria (blog). July 24, 2019. https://zephoria.com/top-15-valuable-facebook-statistics/

Oboler A (2016) Measuring the hate: the state of antisemitism in social media. Online Hate Prevention Institute, Melbourne

O'Callaghan D, Greene D, Conway M, Carthy J, Cunningham P (2015) Down the (White) rabbit hole: the extreme right and online recommender systems. Soc Sci Comput Rev 33(4):459–478

Pacheco E, Melhuish N (2020) Online hate speech: a survey on personal experiences and exposure among adult New Zealanders. eSafety Research, Sydney

Pavlopoulos J, Malakasiotis P, Androutsopoulos I (2017) Deeper attention to abusive user content moderation. In: Proceedings of the 2017 conference on empirical methods in natural language processing. Association for Computational Linguistics, Copenhagen, Denmark, pp. 1125–1135

Perrigo B (2019) Facebook says it's removing more hate speech than ever before. But there's a catch. Time, November 27, 2019. https://time.com/5739688/facebook-hate-speech-languages/

Polymatter (2016) Why YouTube comments suck (and Reddit comments don't). December 15, 2016. https://www.youtube.com/watch?v=Lvf8koqX_yE

Regnér L (2014) The YouTube-born terrorist. J Exit-Deutschland. Z Deradikalisierung Demokratische Kultur 2(0):139–89

Ribeiro MH, Ottoni R, West R, Almeida VAF, Meira W (2020) Auditing Radicalization Pathways on YouTube. In Hildebrandt M, Castillo C (Eds), Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. (pp. 131–141). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372879

Roose K (2019) The making of a YouTube radical. N Y Times, June 8, 2019. https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

Rose B (2013) YouTube comments will soon be less racist, homophobic and confusing. Gizmodo Australia. September 25, 2013. https://www.gizmodo.com.au/2013/09/youtube-comments-will-soon-be-less-racist-homophobic-and-confusing/

Rose-Stockwell T (2018) Facebook's problems can be solved with design. Quartz. April 30, 2018. https://qz.com/1264547/facebooks-problems-can-be-solved-with-design/

Sadowski J (2019) When data is capital: datafication, accumulation, and extraction. Big Data Soc 6(1):2053951718820549

Saima S (2019) YouTube boasts 2 billion monthly active users, 250 million hours watched on TV screens every day. Digital Information World (blog). May 4, 2019. https://www.digitalinformationworld.com/2019/05/youtube-2-billion-monthly-viewers-250-million-hours-tv-screen-watch-time-hours.html

Schmitt JB, Rieger D, Rutkowski O, Ernst J (2018) Counter-messages as prevention or promotion of extremism?! The potential role of YouTube. J Commun 68 (4):780–808. https://doi.org/10.1093/joc/jqy029

Silverstein J (2018) Robert Bowers, Pittsburgh shooting suspect, was avid poster of anti-semitic content on gab. CBS News. October 28, 2018. https://www.cbsnews.com/news/robert-bowers-gab-pittsburgh-shooting-suspect-today-live-updates-2018-10-27/

Solsman J (2018) YouTube's AI is the puppet master over most of what you watch. CNET. January 10, 2018. https://www.cnet.com/news/youtube-ces-2018-neal-mohan/

Statt N (2018) Undercover Facebook moderator was instructed not to remove fringe groups or hate speech. The Verge. July 17, 2018. https://www.theverge.com/2018/7/17/17582152/facebook-channel-4-undercover-investigation-content-moderation

Stevenson A (2018) Facebook admits it was used to incite violence in Myanmar. N Y Times, November 6, 2018, sec. Technology. https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

Tait A (2016) Why are YouTube comments the worst on the Internet? New Statesman, October 26, 2016. https://www.newstatesman.com/science-tech/internet/2016/10/why-are-youtube-comments-worst-internet

Tufekci Z (2018) YouTube, the great radicalizer. N Y Times, June 8, 2018. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Tulkens S, Hilte L, Lodewyckx E, Verhoeven B, Daelemans W (2016) The automated detection of racist discourse in dutch social media. Comput Linguist Netherlands J 6:3–20

Tura N, Kutvonen A, Ritala P (2018) Platform design framework: conceptualisation and application. Technol Anal Strategic Manag 30(8):881–894. https://doi.org/10.1080/09537325.2017.1390220

Vincent J (2017) Former Facebook Exec says social media is ripping apart society. The Verge, December 11, 2017. https://www.theverge.com/2017/12/11/16761016/former-facebook-exec-ripping-apart-society

Vincent J (2019) AI won't relieve the misery of Facebook's human moderators. The Verge. February 27, 2019. https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–51. https://doi.org/10.1126/science.aap9559

Wallaroo Media (2019) Facebook news feed algorithm history. Wallaroo Media (blog). July 3, 2019. https://wallaroomedia.com/facebook-newsfeed-algorithm-history/

Williams ML, Burnap P, Javed A, Liu H, Ozalp S (2020) Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. British J Criminol 60(1):93–117. https://doi.org/10.1093/bjc/azz049

Yablonski J (2019) Humane by design. 2019. https://humanebydesign.com

Zannettou S, Finkelstein J, Bradlyn B, Blackburn J (2020) A quantitative approach to understanding online antisemitism. In De Choudhury M (Ed.), Proceedings of the International AAAI Conference on Web and Social Media (Vol. 14, pp. 786–797). https://www.aaai.org/ojs/index.php/ICWSM/article/view/7343

Zuckerberg M (2018) A blueprint for content governance and enforcement. Facebook. November 15, 2018. https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

Acknowledgements

This article drew extensively upon work from "Angry by Design," a research project that was partially funded by a grant from Netsafe New Zealand.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.M.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.